USING THE DRAFT H.26L VIDEO CODING STANDARD FOR MOBILE APPLICATIONS

Gary J. Sullivan

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
United States of America
garysull@microsoft.com

Thomas Wiegand

Heinrich-Hertz-Institute
Image Processing Department
Einsteinufer 3710587 Berlin
Germany
wiegand@hhi.de

Thomas Stockhammer

Inst. for Communications Engineering
Munich University of Technology
80290 Munich
Germany
stockhammer@ei.tum.de

ABSTRACT

The main goals of the new ITU-T H.26L standardization effort are enhanced compression performance and provision of a "network-friendly" packet-based video representation addressing "conversational" (i.e., video telephony) and "non-conversational" (i.e., storage, broadcast, or streaming) applications. Hence, the H.26L design covers a Video Coding Layer (VCL), which provides the core high-compression representation of the video picture content, and a Network Adaptation Layer (NAL), which packages that representation for delivery over a particular type of network. The H.26L VCL test model has achieved a significant improvement in rate-distortion efficiency – providing nearly a factor of two in bit-rate savings when comparing against the H.263+ test model. NAL designs are being developed to transport the coded video data over existing and future networks such as circuit-switched wired networks, IP networks with RTP packetization, and 3G wireless systems. The NAL is also designed to enable simple gateway operation in heterogeneous transmission scenarios.

1. INTRODUCTION

H.26L [1] is the current project of the ITU-T Video Coding Experts Group (VCEG) – a group officially chartered as ITU-T Study Group 16 Question 6. The primary goals of the H.26L project are:

- *Improved coding efficiency.* The syntax of H.26L should permit an average reduction in bit rate by 50% compared to H.263+ (version 2 of H.263 [3]) for a similar degree of encoder optimization. For that comparison, the encoding algorithms that are specified in the test models for H.26L and H.263+ are used. These are TML-7 [1] and TMN-10 [2], respectively, with both using Lagrangian coder control [5].
- *Improved Network Adaptation.* Issues relating to network adaptation that were examined seriously for the first time in the H.263 and MPEG-4 projects [3, 4] are being taken further in H.26L. The scenarios emphasized are primarily for Internet, LAN, and third-generation mobile wireless channels.
- *Simple syntax specification.* The design of H.26L is strongly intended to lead to a simple and clean solution avoiding any excessive quantity of optional features or profile configurations.

While the H.26L design is still a work-in-progress, a draft design was adopted in August 1999 and has evolved into a *test model long-term* (TML) reference design, with TML-7 being the latest version at press time [1]. A new feature of the design is its introduction of a conceptual separation between a Video Coding Layer (VCL), which provides the core high-compression representation of the video picture content, and a Network Adaptation Layer (NAL), which packages that representation for delivery over a particular type of network. Figure 1 shows the overall concept of the H.26L design.

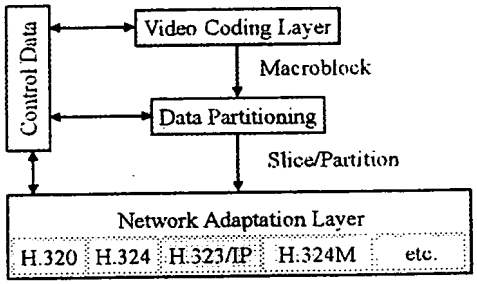

Fig. 1: Block diagram of an H.26L coder consisting of Video Coding Layer (VCL) and Network Adaptation Layer (NAL).

The H.26L draft VCL design is basically similar to that of prior standards, in that it is a block-based motion-compensated hybrid transform video coder. As with prior standards, only the decoding process is precisely specified to enable interoperability, while the processes for capturing, pre-processing, encoding, post-processing, and rendering are all left out of scope to allow flexibility in implementations. However, in contrast with prior standards, H.26L contains a number of new features that enable it to achieve a significant improvement in coding efficiency relative to prior standard designs.

The H.26L draft NAL design provides the ability to customize the format of the VCL data for delivery over a variety of particular networks. Therefore, a unique packet-based interface between the VCL and the NAL is defined. The packetization and appropriate signaling is part of the NAL specification, which is not necessarily part of the H.26L specification itself. For the transmission of video in 3G mobile systems like UMTS and CDMA-2000 with limited bandwidth and transmission power resources, the necessity for high compression efficiency is obvious. However, due to special properties of the wireless channel and mobile systems, an adaptation of the video data to the network's specific properties

is an additional important task. These two design goals, compression efficiency and network friendliness, motivate the differentiation between the VCL for coding efficiency and the NAL to take care of network issues.

2. THE H.26L VIDEO CODING LAYER

The H.26L draft VCL consists of an efficient block-based motion-compensated hybrid transform video coder that contains a number of innovative features.

2.1. H.26L Motion Representation

In H.26L, conventional Intra (I), Inter (P), and Bi-directional (B) picture types are supported. Inter-frame motion is represented by the translational displacement of block-shaped regions from previously-encoded pictures. These block shapes can vary (a feature that first appeared in a more primitive form in Annex F of H.263 version 1 [3]), with the design including blocks of 16×16, 8×16, 16×8, 8×8, 4×8, 8×4, and 4×4 pixels.

The H.26L design supports 1/4 and 1/8 sample motion vector accuracy. Quarter-sample motion first appeared in standards in MPEG-4 version 2 [4], although with different interpolation filtering. In contrast to MPEG-4's block boundary mirroring and 8-tap filtering for half-pel interpolation with quarter-pel motion, H.26L uses 6-taps and makes use of a "special position" with extra low-pass filtering to reduce high-frequency noise. Eighth-sample motion (intended for more specialized advanced-profile applications and new to standardization with H.26L) uses 8-tap filtering [9].

Multi-frame motion-compensated prediction is supported, allowing the encoder to use more than one prior coded picture as a reference for the coding of each additional picture. The long-term memory feature found in H.26L first appeared in standard codecs as Annex U of H.263 version 3 [3].

A number of other new features are included in H.26L motion representation. A new type of picture called an SP picture is defined that enables a predictive switching between different video streams or between different parts of a single video stream [7]. This is accomplished by placing a forward transform and quantization operation in the decoder processing as part of the motion-compensated prediction process. The primary intended application for SP pictures is server-based streaming. Also, P pictures can be predicted from temporally-subsequent reference pictures, B pictures are generalized with an explicit weighting factor for prediction averaging, and motion vectors can cause extrapolation beyond the reference picture boundaries (as first found in Annex D of H.263 version 1 [3]).

2.2. H.26L Transform Processing

H.26L is similar to prior standards in its use of a block transform that is essentially an inverse discrete cosine transform (IDCT). However, there are several significant differences between H.26L's transform design and that of prior standards. The primary difference is that H.26L's transform is primarily a 4×4 in shape. To extend the length of the basis functions for smooth regions an additional 4×4 transform is applied to DC values from sixteen 4×4 blocks of intra luminance data, and a 2×2 transform is applied to DC values for chrominance data. The transform is an exact integer operation rather than a specification in terms of accuracy tolerances for a real-valued transform (thus avoiding inverse-transform mismatch) – a concept that first appeared in Annex W of H.263 version 3 [3], but is taken further in H.26L. Normalization of the coefficient scaling is folded into the inverse quantization process for reduced complexity.

H.26L uses scalar inverse quantization, but without the extra-wide dead-zone found in typical prior designs. For improved rate control capability, quantization step sizes are controlled in approximately 12.5% increments, rather than in increments of fixed size as found in prior standards.

2.3. H.26L Entropy Coding and Coefficient Scanning

H.26L uses innovative entropy coding schemes that simplify the VCL design. Two methods of entropy coding are supported in H.26L. They are the Universal Variable-Length Coder (UVLC) and the Context-Adaptive Binary Arithmetic Coder (CABAC).

The UVLC uses one infinite-extent codeword set. Rather than designing a different code for each element of the H.26L syntax, only the mapping to the single UVLC code table is customized to the probabilistic behavior of the data. Coefficient scanning is done primarily with a run-level 2-D scheme with an end-of-block symbol (as found first in H.261 [6]). For fine quantization, a special "double scan" is substituted to fit coefficient statistics more closely to the UVLC design.

For maximal efficiency, the CABAC technique can be applied after the UVLC [10]. CABAC provides both an ability to adapt to local source statistics and extremely efficient entropy coding capability. Although Annex E of H.263 version 1 [3] also included an arithmetic coding capability for video data, the adaptivity of the CABAC design can provide a more significant improvement in entropy coding performance, especially when very fine or very coarse quantization is in use.

2.4. H.26L Deblocking Filter

The H.26L design includes an in-loop deblocking filter for removing block-edge artifacts. The concept of this filter is basically similar to that specified in Annex J of H.263 version 2 [3], but the rounding error problems of that prior specification are removed by use of the exact inverse transform.

2.5. H.26L Intra Prediction

Intra pictures are coded with non-temporal prediction of the picture representation. The use of prediction within the picture has appeared before (e.g. prediction of DC and AC coefficient values in Annex I of H.263 version 2 and in MPEG-4 version 1 [3, 4]), but greater efficiency is achieved in H.26L by use of directional prediction in the spatial domain rather than coefficient value prediction in the transform domain.

2.6. H.26L Coding Efficiency Experiments

Experiment results are provided in Figure 2 to show H.26L coding efficiency. The four cases compared herein are:

- **H.263: FP-MC**: TMN-10 encoding, but with only "baseline" features and full-pel accuracy motion compensation used in H.263 standard syntax. This case roughly corresponds to H.261 [6] performance.

- **H.263: Baseline**: TMN-10 encoding, but only baseline syntax features are permitted for the H.263 standard [3].
- **H.263: TMN-10**: The full TMN-10 coder [2] using Annexes D and F of H.263 version 1 and Annexes I, J, and T of H.263+ [3]. This is used as a reference for the 50% rate reduction goal.
- **H.26L: TML-7**. The TML-7 coder [1] (per TML 6.2 software) with all TML-6 features and CABAC enabled.

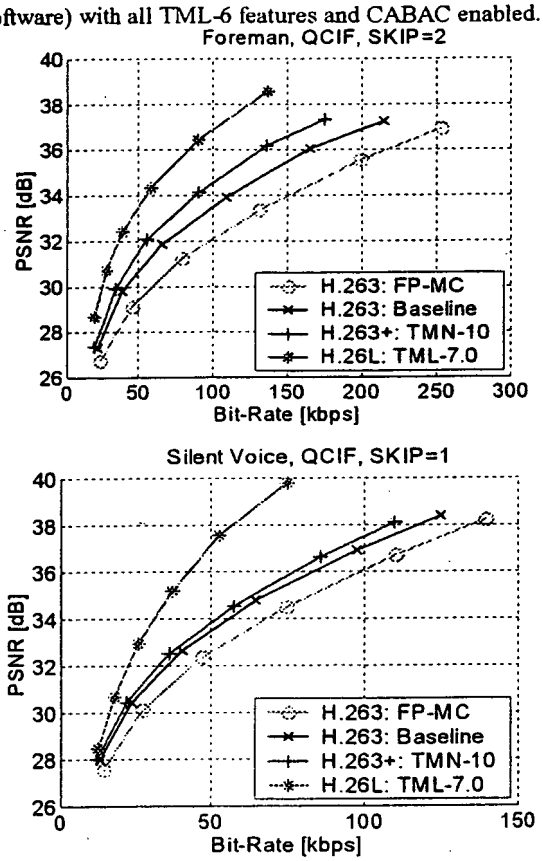

Fig. 2: H.26L coding results. In all experiments, H.26L's TML-7 clearly outperforms H.263's TMN-10. Significant improvement has been found for a wide variety of sequence characteristics, picture resolutions, and bit rates.. TML-7 provides roughly a 40% bit-rate reduction against TMN-10.

4. H.26L FUNCTIONALITIES SUPPORTING MOBILE VIDEO APPLICATIONS

3.1 Video Applications on Mobile Systems

Video transmission for mobile terminals will be a major application in the upcoming 3G systems and may be a key factor in their success. The display of video on mobile devices opens the road to several new applications. Three major service categories are identified in the H.26L standardization process [8]: *1)* conversational services for video telephony and video conferencing, *2)* live or pre-recorded video streaming services, and *3)* video in multimedia messaging services (MMS).

In general, mobile devices are hand-held and constrained in processing power and storage capacity. Therefore, a mobile video codec design must minimize terminal complexity while remaining consistent with the efficiency and robustness goals of the design. In addition, the mobile environment is characterized by harsh transmission conditions in terms of fading and multi-user interference, which results in time- and location-varying channel conditions. Many highly sophisticated radio link features like broadband access, diversity techniques, fast power control, interleaving, forward error correction by Turbo codes, etc., are used in 3G systems to reduce the channel variance and therefore the bit error rate and radio block loss rate.

Entirely error-free transmission of radio blocks is a generally unrealistic assumption. – although with RLC retransmission methods, delay insensitive applications like MMS can be delivered error-free to the mobile user. In contrast, conversational and streaming services with real-time delay and jitter constraints allow for only a very limited number of retransmissions, if any [11]. In addition, in a cellular multiuser environment the transmission capacity within each cell is limited. Therefore if new users enter the cell, active users must share resources with new users and if users exit the cell, the resources can be re-allocated to the remaining users. The well-designed 3G air interfaces allow data rate switching in a very flexible way by assigning appropriate scrambling or channel coding rates. This results in a need for video codecs to be capable of (for example) doubling or halving video data rate every 5-20 seconds. Therefore, due to the time-varying nature of the mobile channel, the video application must be capable of reacting to variable bit-rate (VBR) channels as well as to residual packet losses. Finally, the prioritization and quality of service design for mobile links is an ongoing standards and research activity. Systems supporting prioritized transmission need video codecs generating data with different priorities.

Finally, for network-specific issues like encapsulation, signaling of setup and control information, synchronization or feedback schemes, the mechanisms of the underlying transport layer should be used as efficiently as possible. To summarize this discussion, the following (probably incomplete) list of functionalities are required of a video codec to operate on 3G mobile devices:

- Very high compression efficiency,
- Support of low power, low memory, low complexity decoding and, for some applications, encoding,
- Robustness to packet losses,
- Support of short-term small-range variable bit-rate channels,
- Support of online long-term large range rate-switching mechanism,
- Generation of different priority classes,
- Efficient and appropriate usage of network specific mechanisms.

3.2 Error Resilience and Functionality Tools

The previous discussion motivates several tools included in the H.26L draft for error resilience and enhanced functionalities to support the transmission of different video applications over 3G mobile systems. In addition to high compression efficiency, the following tools are part of the H.26L draft.

For enhanced error resilience the test model allows to interrupt spatial, temporal and syntactical predictive coding. The principles of each of the adopted features are reasonably well known from prior video coding work, particularly from the H.263+ and MPEG-4 projects [3, 4]. However, these features are taken a bit further in the H.26L design. Temporal resynchronization within an H.26L video bitstream can be accomplished by use of intra picture refresh (stopping all prediction of data from one picture to another), whereas spatial resynchronization is supported by slice structured coding (providing spatially-distinct resynchronization points within the video data for a single picture). In addition, the usage of intra macroblock refresh and multiple reference frames allows the encoder to decide the macroblock mode not just on compression efficiency criteria but also by taking into account the loss characteristics of the transmission channel [12, 13].

Fast rate adaptation can be accomplished by switching the quantization fidelity on a macroblock basis such that a real-time encoder can react immediately to varying bit rate. For streaming of pre-coded sequences, well-designed buffering can deal reasonably well with varying bit-rate conditions. Still, buffer overflows in VBR environments may not be completely avoidable. The introduction of a syntax-based data partitioning scheme with multiple partitions per frame can allow less important information to be dropped in the event of a buffer overflow. With the use of multiple reference pictures in P- and B-frames, similar policies might be applied using temporal data partitioning. Note that data partitioning concepts can also be used to generate video data with different priority classes to support quality of service concepts in networks.

As outlined, in addition to small range bit-rate variations especially prevalent in mobile environments, rapid switching of data rates in larger ranges is a desirable feature. By the possibility to change quantization fidelity as well as spatial and temporal resolution for each and every frame for real-time encoding, large bit-rate variations are supported. However, for the highly-prevalent case when pre-coded sequences or multicast streaming are used, encoder reaction to the varying bit rates is impossible. Fine Granular Scalability (FGS) has recently been adopted into MPEG-4 to deal with these effects [4, 14]. However, the low efficiency of that current FGS approach motivates the use of stream switching instead of scalable coding. In the use of current video coding standards it is common to include periodic I-frame refreshes to allow the switching. H.26L defines a new picture type, called an SP-frame [7], to allow switching between versions of a stream without introducing the efficiency loss associated with an I-frame.

3.3 NAL Design for Mobile Applications

In addition to the functionalities presented above, the integration of the video data into the network specific transport format is an additional new concept in the test model. The Network Adaptation Layer (NAL) is responsible for encapsulation the data provided in the slice format using the specific properties of the underlying networks. This includes, for example, framing, signaling of logical channels, usage of timing information or end of sequence signaling.

Among others, an NAL will be specified for transmission over H.324/M for circuit switched conversational services in 3G systems as well as an RTP/UDP/IP NAL format to support video over IP. The only VCL structure to be accessed by the NAL is the slice structure containing a header and payload data. The payload might consist of several partitions, if data partitioning is applied. The header contains all relevant information for a particular slice, and the task of the NAL is to provide an appropriate mapping of the header and payload information onto the transport protocol (e.g. framing and resynchronization overhead like picture start codes can be avoided in packet switched transmission). Specification of how to transmit setup and sequence control information as well as other network specific tasks for several networks will be specified in the NAL.

CONCLUSIONS

The H.26L project promises some significant advances in the state of the design art in standardized video coding, including key aspects designed with mobile applications in mind. Further information, documents and software for the project can currently be found at http://standard.pictel.com/ftp/video-site and http://kbs.cs.tu-berlin.de/~stewe/vceg/.

REFERENCES

[1] ITU-T/SG 16/VCEG (formerly Q.15 now Q.6), H.26L Test Model Long-Term Number 7 (TML-7), Doc. VCEG-M81, Apr. 2001.

[2] ITU-T/SG 16/VCEG (formerly Q.15, now Q.6), Video Codec Test Model Near-Term Number 10 (TMN-10), Tampere, Apr. 1998.

[3] ITU-T, Video Coding for Low Bit-Rate Communication, ITU-T Recommendation H.263, Version 1: November 1995, Version 2: January 1998, Version 3: Nov. 2000.

[4] ISO/IEC JTC1, Generic Coding of Audiovisual Objects – Part 2: Visual (MPEG-4 Visual), ISO/IEC 14496-2, Version 1: January 1999, Version 2: January 2000; Version 3: January 2001.

[5] G. J. Sullivan and T. Wiegand "Rate-Distortion Optimization for Video Compression," in *IEEE Signal Proc. Magazine*, vol. 15, no. 6, pp. 74-90, Nov. 1998.

[6] ITU-T, Video Codec for Audiovisual Services at px64 kbit/s, ITU-T Recommendation H.261, Version 1: Nov. 1990; Version 2: Mar. 1993.

[7] R. Kurceren, M. Karczewicz, "A Proposal for SP-frames," ITU-T SG 16 Doc. VCEG-L27, Eibsee, Germany, Jan. 2001.

[8] S. Wenger, M. Hannuksela, T. Stockhammer, "Identified H.26L Applications," ITU-T SG 16, Doc. VCEG-L34, Eibsee, Germany, Jan. 2001.

[9] Thomas Wedi, "1/8-pel Displacement Vector Resolution for TML-6," ITU-T SG16 Doc. VCEG-M45, Austin, TX, USA, Apr. 2001.

[10] D. Marpe, G. Blättermann, G. Heising und T. Wiegand, "Further Results for CABAC entropy coding scheme," ITU-T SG16 Doc. VCEG-M59, Austin, TX, USA, Apr.

2001.

[11] G. Roth, R. Sjöberg, G. Liebl, T. Stockhammer, V. Varsa, and M. Karczewicz, "Common Test Conditions for RTP/IP over 3GPP/3GPP2," ITU-T SG16 Doc. VCEG-M77, Austin, TX, USA, Apr. 2001.

[12] R. Zhang, S.L. Regunathan, and K. Rose, "Video Coding with Optimal Inter/Intra-Mode Switching for Packet Loss Resilience," in IEEE JSAC, vol. 18, no. 6, pp. 966-976.

[13] T. Wiegand, N. Färber, B. Girod, "Error-Resilient Video Transmission Using Long-Term Memory Motion-Compensated Prediction," in IEEE JSAC, vol. 18, no. 6, pp. 1050-1062, June 2000.

[14] W. Li, "Overview of Fine Granular Scalability in MPEG-4 Video Standard," IEEE Trans. On CSVT, vol. 11, no. 3, pp. 385-398, March 2001.